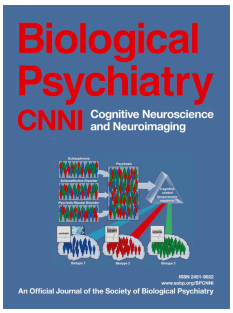
DOI:
10.1016/j.bpsc.2018.06.010

Document Version
Peer reviewed version

Accepted Manuscript



Active Inference on OpenAI Gym: A Paradigm for Computational Investigations into Psychiatric Illness

Maell Cullen, Ben Davey, Karl J. Friston, Rosalyn J. Moran

PII: S2451-9022(18)30161-7

DOI: 10.1016/j.bpsc.2018.06.010

Reference: BPSC 301

To appear in: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*

Received Date: 4 April 2018

Revised Date: 23 June 2018

Accepted Date: 25 June 2018

Please cite this article as: Cullen M., Davey B., Friston K.J. & Moran R.J., Active Inference on OpenAI Gym: A Paradigm for Computational Investigations into Psychiatric Illness, *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (2018), doi: 10.1016/j.bpsc.2018.06.010.

Active Inference on OpenAI Gym: A Paradigm for Computational Investigations into Psychiatric Illness

Maell Cullen[1+*], Ben Davey[1+], Karl J. Friston[2], Rosalyn J. Moran[1,3]

[1*]Department of Engineering Mathematics, Merchant Venturers School of Engineering, University of Bristol, 75 Woodland Rd, Bristol BS8 1UB, UK.

[2]Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, London WC1N 3BG, UK

[3]Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, London, SE5 8AF, UK.

Abstract word count: 248

Main text word count: 3981

Number of Tables: zero

Number of Figures: 7

Number of Supplementary Materials: 1

Keywords : active inference; game-based imaging biomarkers; markov decision process; free energy principle; computational psychiatry; computational phenotyping

[+] Equal Contributions

*To whom correspondence should be addressed:

Maell Cullen,

Department of Engineering Mathematics,

Merchant Venturers School of Engineering,

University of Bristol,

75 Woodland Rd,

Bristol, BS8 1UB,

UK

maell.cullen@bristol.ac.uk

Background:

Artificial Intelligence has recently attained human-like performance in a number of 'game-like' domains. These advances have been spurred by brain-inspired architectures and algorithms such as hierarchical filtering and reinforcement learning. OpenAI Gym is an open-source platform to train, test and benchmark algorithms – providing a range of tasks including classic arcade games such as DOOM. Here we describe how the platform might be used as a simulation, test and diagnostic paradigm for psychiatric conditions.

Methods:

To illustrate how Active Inference models of game-play could be used to test mechanistic and algorithmic properties of psychiatric disorders we provide two exemplar analyses. The first speaks to the impact of ageing on cognition, examining game play behaviours in a model of ageing where we compare age-dependent changes of younger (n=9, 22 ± 1 years) and older (n=7, 56 ± 5 years) adult players. The second is an illustration of a putative feature of anhedonia – where we simulate diminished sensitivity to reward.

Results:

These simulations demonstrate how Active Inference can be used to test predicted changes in both neurobiology and beliefs in psychiatric cohorts. As well as behavioural measures we show that putative neural correlates of Active Inference can be simulated and hypothesized (model-based) differences in local field potentials and BOLD responses produced.

Conclusions:

We show that Active Inference, through epistemic and value-based goals, enables simulated subjects to actively develop detailed representations of gaming environments and demonstrate the usage of a principled algorithmic and neurobiological framework for testing hypotheses in psychiatric illness.

Introduction

Recent perspectives on psychiatric illness highlight the crucial role of computational assays in deciphering the complexities of mental illness (1,2). Computational assays of the cognitive and behavioural abnormalities that arise from psychiatric illness provide a formal mapping from complex thought disorders to putative neural substrates (3). The central proposition here is that cognition and behaviour are emergent features of biological processes and that by capturing these processes formally, we may better access the origins of psychiatric disease. Several computational frameworks have been deployed recently for the purposes of understanding neuropsychiatric diseases, including Bayesian learning (4), drift diffusions processes (5) and temporal difference models (6).

Neuroimaging advances have provided useful biological insights into regional and connectivity deficits associated with mental illness (7) although these methods require further development to translate into pragmatic clinical tools for diagnostic and prognostic classifications at the individual level (8). Mathematical models that provide both individual brain and behavioural predictions are potentially even more powerful candidates to advance the field of computational psychiatry as they provide both algorithmic (information-processing) and biophysical insights that link psychopathology and pathophysiology within a single framework (2).

Reinforcement learning (RL) models with model-based fMRI (9) have predominated in this area, with neuromodulators and decision-making circuits in the striatum (10) highlighted as crucial neural substrates linking aberrant decision making and learning to psychiatric disease symptoms. Using this technique, simulated events, based upon an RL model, are convolved with a hemodynamic response function and correlated against BOLD fMRI signals to associate decision-making processes with the brain regions in which they originate (11). Here the aim is to produce regressors with powerful explanatory capabilities in terms of adaptive

reward-seeking and punishment-avoiding behaviours. The appeal of RL models lies predominantly in the expression of *valence* prediction errors and their association with mental disorders such as anhedonia in depression (12) and impulsivity in ADHD (13). Similarly, hierarchal Bayesian analysis has been used in conjunction with model-based fMRI to model prediction errors (14). Here, the focus is not specifically on reward or punishment prediction errors but errors in relation to prior beliefs about states in the world more generally. For example, Ahn et al (15) corrrelate choice behaviours with decision-time activation in the ventromedial prefrontal cortex. When compared with non-Bayesian methods, this technique was found to provide more accurate individual and group estimates and modelling predictions.

The (Bayesian) formalism provided by the free energy principle and active inference (16) stipulates that the brain should maintain a model of the world that can predict incoming signals from the environment. A model that can minimize long-term surprise by making accurate predictions, will thus have high 'model evidence'. In the current setting, resolving the moment to moment free energy of the brain, ensures the minimisation of long term surprise. The free energy in turn is made up of the differences between the model predictions and the sensed data from the environment (prediction errors) which algorithmically are scaled by the certainty of that prediction (precision-weighted prediction errors) (17). The free energy principle thus appeals to the dual goals of computational psychiatry as the neurobiological circuits required for this form of 'active inference' overlap with key anatomical features; e.g., precision-weighted prediction errors transmitted through cortical microcircuitry in the sensory cortices (18, 19).

As introduced in (20) computational psychiatric approaches could be usefully understood by considering normative models and process models of the brain. Normative models refer to the mathematical or computational goal of the brain or agent – without regard for how it might be

implemented in the brain (e.g. learning), while process models refer to the mechanisms that might implement a particular algorithm (e.g. synaptic plasticity via LTP). Active inference accounts of brain function purport to do both. It comprises a general algorithm which might be the goal of brains (to minimize free energy) while also proposing distinct neurobiological components to implement the running of free energy minimization via distinct message passing sequences between prefrontal, sensory and neuromodulatory neurons (16).

Importantly, elements of the belief update procedure within this modelling framework are constrained by variational approximations to Bayesian inference that have been mapped to putative neurobiological components or processes (21). However, it remains to be tested whether these distinct neurobiological processes do align with model inversion dynamics or whether such components of inference generalize from one task to another.

The algorithmic framework afforded by the free energy principle may also be applied at an individual level, to characterize individuals with respect to their prior beliefs and preferences by fitting their choice behaviours to a computational model (22). Thus, given a sufficiently simple design for patient populations, where the task structure can be formalised, distinctions between cortical and subcortical effects on behaviour can be distinguished in the usual manner of using orthogonalized regressors (9)**.** This framework deconstructs pathological behaviour with respect to individual's generative model of a given task under the assumption that the parameterisation this model may be used to predict their behaviour under various cognitive protocols. This is pertinent to computational psychiatry in that features of psychopathology that are likely to arise from dysfunctional models of the environment, maladaptive learning or failures of inference (22) may be evaluated across different tasks and laboratories.

This paper is concerned with constructing and demonstrating the use of generative probabilistic models that can explain psychopathology – under the free energy formalism – to produce behavioural and imaging features that can be tested empirically at an individual level within a game environment. We use the OpenAI Gym platform (23) as it provides standardized tasks and computational environments, allowing for comparative models of behaviour to be shared across the reinforcement learning community. By altering the parameters of the generative model and therefore of the inference procedure, we aim to demonstrate with two toy examples how decision making may be altered in an identifiable way through neurobiological changes associated with psychiatric illness or ageing.

Methods and Materials

The DOOM Environment on OpenAI Gym

DOOM is a well-known pseudo-3D game that has been used as a platform for reinforcement learning (24) and computer vision (25). DOOM was chosen to demonstrate the versatility and potential for gaming environments and platforms such as Gym for computational psychiatry. It also provides a simple game scenario for comparisons of free energy and reward-maximising schemes. To prepare the game for active inference experiments we construct three variables – the A matrix, B matrix and C vector. The A matrix comprises the agent's belief about the mapping between sensory information and states of the environment; this mapping is simplified by deconvolving the pixel data into a set of manageable corner features using the Harris corner detection algorithm (42). Any suitable technique may be used to achieve this goal; however, within this environment it is reasonable to assume that the target will exhibit the largest number of corner features within the visual scene. This implies that the current state (position of the target relative to the player) can be defined as the location exhibiting the greatest response to the Harris corner operator. The B matrix comprises the agent's beliefs about possible transitions between states under actions while the C vector comprises beliefs about expected outcomes, serving as a proxy for utility or reward. A detailed description of gameplay and the DOOM state space is provided in the supplementary material while feature extraction methods are illustrated in Figure 1.

Human Play of DOOM

To assess whether our simulated agents could attain 'human-like' performance we collated scores from a sample of real players, playing the same DOOM game. We measured the performance of 16 players, 8 female aged 37 ± 17 years (mean ± std). From these results we formed two sets of player data, comprising younger (n=9, 22 ± 1 years) and older (n=7, 56 ±

5 years) adult players for comparison with our 'young' (10-state) and 'older' (6-state, see supplementary materials) agents. These simulations were designed to represent and contrast simpler (older) generative models of the world with more complex (younger) models of the 'DOOM' world.

Manipulations to simulate features of Anhedonia

Anhedonia is a behavioural trait of individuals with depression characterized by a lack interest in rewarding or pleasurable activities (26, 27). The anhedonic aspect of depression has been previously examined in a large reinforcement learning meta-analysis comparing learning mechanisms with reward sensitivity (12). There they found that among two alternative hypotheses of anhedonic responses (abnormal learning vs. diminished reward sensitivity), that diminished reward sensitivity most parsimoniously explained a large literature on reward-based reinforcement paradigms in depression. Under active inference, diminished reward sensitivity can be represented by adjusting the prior beliefs about expected outcomes (C vector) to reflect a less optimistic view of the world and the 'winning' state in the game of DOOM. This selection of prior beliefs in an 'anhedonic' agent effectively means that the agent is indifferent to any particular outcome. The comparative analysis thus comprises a 'motivated agent' and an 'anhedonic agent' with identical knowledge of the environmental structure, but different prior beliefs about the outcomes or 'goals', illustrated in Figure 7A. Due to the form of the prior belief vectors, we can say that both agents know, a priori, 'where' the reward is, but the scale of the reward is reduced for the anhedonic agent. Our simulations would thus include the transitive effect of diminished reward sensitivity on learning – which is influenced by the states that are visited by the agent – and decisions to visit these states, which are driven by prior beliefs.

Simulating Neural Responses

To generate neurobiological predictions related to symptoms of anhedonia and potential imaging biomarkers for depression that may be observable during gameplay we used the simulations from the agents above and analysed the belief updates over trials across 128 episodes. This requires policies as well as past and potential future states to be 'kept in mind'. Hence, these belief updates under the variational scheme may be represented in online, working memory areas such as the prefrontal cortex. We sought to test how alterations in goal states or expected beliefs about outcomes would alter state estimation and the prefrontal responses that could subtend this inference under a 'motivated' and 'anhedonic' set of goals. In our results we illustrate putative local field potentials (LFPs) and BOLD responses within the PFC and provide a guide for generating these simulated responses within the supplementary material.

Results

Active Inference Builds More Complete Representations of DOOM

Figure 2 illustrates how the simulated free energy minimizing agent plays the game DOOM and how it compares to a classic reward (goal) maximizing scheme (Eqn. 1 in Supplementary Material). In these simulations we report the reward and survival metrics returned by Gym over each episode. We assumed that the agent holds a 6-state representation of the game (Figure 3A). The agents began at episode 1 with the assumption that states map directly to outcomes, that selecting an action of 'move left', 'move right' or 'fire' will result in moving from any state to any other state with equal probability (equal entries of 1/6 in the *B* matrices) and that the desired state is state 4 (in the middle firing), with firing outside of range the least desirable states. Only policy selection depended on whether we used epistemic and extrinsic value (free energy minimizing) or only the extrinsic value (reward maximising, Eqn. 1 in supplementary material) to guide behaviour.

Figure 2 shows one such instance from an agent that is driven by the imperative to minimize free energy and an agent driven by the imperative to maximise reward. Figure 2B illustrates the *B* matrices of each agent at timesteps 4, 16, 64, and 128. This serves as a visualisation of the agent's emerging understanding of state-action dynamics as the they move through the environment. Interestingly, around trial 80 (Figure 2C), the agent's performance begins to decline. This is due to the agent learning an incorrect state transition; most likely due to a failure of the Harris Corner detection algorithm (Figure 1). However, by t=128 the agent has relearned a full and correct representation of the environment, overcoming the earlier erroneous state transitions (Figure 2B). The learned state transitions under reward maximization are demonstrably less robust. Figure 2B (lower) shows that the agent has learned little about the causal structure of the environment, indicated by uniformity of the transition matrices. The agent has a lower average reward due to an inability to form an

optimal policy for navigating the environment (Figure 4). Here reward 'scores' are significantly lower over instances for reward-based decision making as compared to free energy-based decision making, $p = 0.05$ and for 'survival' scores where free energy agents complete the games earlier compared to reward maximizing agents; $p = 0.04$.

Our simulations and illustrations of learning demonstrate that active inference outperforms reward maximizing policies. This is due to a difference in their respective cost functions. Active inference entails the simultaneous maximisation of two components: the epistemic value of an action (reducing uncertainty about state transitions) and the extrinsic value of an action. Under a reward maximizing policy only the latter is optimized. Hence the agent cannot achieve desired goal states with the same effect – since the agent is not driven to learn but only does so through trial and error, resulting in the decreased performance overall (Figure 4).

Active Inference vs. Humans in DOOM

To assess the ability of a free energy minimizing agent to 'compete' with a human player we enrolled 16 participants to play the game. In Figure 5, we show the survival scores from these games and compare them with the free energy agent, taking every other episode from the 128 simulations above. We found that during the very early trials, the free energy agent is exploring and learning the structure of the environment before engaging in exploitative behaviour. This, alongside learning and then unlearning (Figure 2C) maladaptive behaviours, explains the slower transition to behaving optimally. However, matching human performance was remarkably fast, with the free energy agent attaining human-like performance after only 12 actions (Figure 5). This indistinguishable performance was also retained throughout all remaining trials ($p > 0.05$).

Complex and Simple Models of the DOOM world and Ageing

Given the importance of development and ageing in psychiatric disease onset (28) and recovery (29), as well as the notion of model simplification with ageing (30), we asked whether free energy agents can mimic age-dependent play in our human player cohort. We first described two free energy models that represent complex and simple models (Figure 3). We show that the simple model performs well (and equally well after 12 decisions) when compared to a human agent.

From our human sample we compared play from 9 younger (37 ± 17 years old) with 7 older (56 ± 5 years old) participants. To quantify the survival metrics obtained from the young and old participants we fit a quadratic polynomial of order two to these data. We found that over the course of 64 games, the survival metrics of both 6-state and 10-state models (Figure 6A) and to the older and younger human participants (Figure 6B) share features of change. Specifically, we find enhanced negative linear coefficients for the young compared to old human curves (young = -2.3, older= -1.5) which is recapitulated by the difference between 10 state and 6 state agents (10 state = -4.1; 6-state = - 3.5). We also find that the second order quadratic curvature of these average survival curves are greater for the younger compared to older player games (young = 0.029; older = 0.016) – which again are reflected in the agents (10 state = 0.024; 6 state = 0.021). Overall this might reflect similarities in terms of learning efficiencies between the younger compared to older players and the more complex compared to more simple state-representations.

Simulating Game Play under Anhedonic Priors

To simulate features of depression in simulated play, we developed a new agent whose belief in final outcomes was relatively flat (Figure 7A). In contrast, our 'healthy' or 'motivated' agent retained similar preferences to our previous simulations; believing that it would end up in front of the target, shooting it (Figure 7A). From 4 simulations of each agent over 64 episodes, we found that on average, the motivated agent outperformed the anhedonic agent ($p = 0.04$). However, interestingly the anhedonic agent still learned the structure of the environment and sought out wins in later trials, indicating intact learning (data not shown).

For neuroimaging predictions, we simulated putative neural correlates of activity in the 'prefrontal cortex' of the anhedonic and motivated simulated players. We found that the amplitude of LFPs from the prefrontal cortex demonstrated a particular temporal excursion in motivated compared to anhedonic agents (Figure 7B). Overall LFPs had similar patterns within and over trials, with triphasic potentials for both anhedonic and motivated agents. Importantly this triplet reduced in amplitude for later potentials at a discrete point of learning over the episodes (Figure 7B). Crucially, the anhedonic agents displayed this qualitative change earlier in the learning episodes. Thus, the difference potentials exhibit an excursion around episode 20, with motivated agents retaining the larger potential triplet for a further 5 trials (Figure 7B). This side-by-side comparison of two agents with different belief structures was replicated in 3 further exemplars, suggesting a consistent alteration in state inference strategy and a concomitant change in LFPs that can be systematically predicted and verified; e.g. using a time-frequency analysis of EEG or MEG for a particular player based upon a set of beliefs and behaviour. We then used these LFPs to generate BOLD responses within the PFC. Here the excursion is also marked, with the BOLD response exhibiting a second small peak at around 22 seconds after the beginning of the game, consistent with the timing of when the LFP response exhibits its qualitative change (Figure 7C). Overall, we can compare individuals in terms of their neural responses for alternative beliefs or goals, while

recapitulating similar forms at the group level; i.e. over different instances of these healthy and pathological agents (Supplemental Figure S3).

Discussion

Here, we present a treatment of game play using a benchmarking framework – OpenAI Gym – and simulate changes in behaviour and brain responses associated with features of healthy neurological ageing and neuropsychiatric disease. Our simulations are based on the theory that living creatures, including humans, seek to minimise free energy (31). Importantly, both neurobiological and algorithmic components are interpretable within this framework and so alterations in abstract cognitive constructs such as 'belief' can be mapped to their putative neural substrates. This is important for modern computational psychiatry where a key assumption driving many computational deconstructions is that psychopathologies such as depression or addiction are likely to arrive from maladaptive alterations in neural circuitry which then subtend dysfunctional models of the environment, maladaptive learning or failures of inference (32, 33, 1).

From our simulations, we find first that unlike an artificial agent that simply seeks to maximise reward, a free energy minimizing agent can develop an internal model of its environment (Figure 2). This suggests that superficial learning linked to the prospect of reward may not be sufficient for building models that serve as analogues for realistic human behaviour. We also compared our simulations to real human players. Although the performance of humans in this toy scenario will be trivial, it is important that the performance of free energy-minimizing (and reward maximizing) agents can be put into context in standardized environments to assess whether it is, in general, fit-for-purpose. Less trivial environments where optimal strategies are unknown or difficult to infer and learn should be used to test human performance. It should be noted that the Active Inference formalism can

be applied in a similar fashion to any paradigm where an agent within a closed environment must perceive, act and make decisions.

Nevertheless, even in a toy task we can identify learning trends that may be reflected in ageing (Figure 6) and demonstrate how computational phenotypes and neural biomarkers may be elucidated from gameplay, with a focus on anhedonic features of depression (Figure 7). A more detailed analysis may consider reaction time and other mechanisms of biological feedback such as eye tracking.

It is not a trivial task to divide a domain, at any level of abstraction into discrete states *a-priori,* but MDPs can be combined with continuous space and time generative models (34), which may offer scope for future active inference applications. We have manually discretised the game environments used in this work but the discretisation of internal representations through perceptual inference may also be the subject of future work.

Games can provide a basis for testing memory, reasoning, sensory-motor capabilities and attention in individuals regardless of their physical and cognitive abilities or their age, gender or culture. The emergent nature of game play thus presents a constrained complexity that can be used to understand interactions between various determinants of an individual's behaviour. Comparing neuroimaging data with behavioural models rather than behaviour allows us to deconvolve complex psychiatric phenotypes that may be used as proxies for the hidden causes that drive aberrant behaviours; where a cause may conceptually encompass anything from an individual's prior beliefs and experiences to their neurobiological idiosyncrasies. In stroke, substance abuse and general-purpose motor rehabilitation, (35, 36, 37) game environments have been employed for recovery. Efforts are also being made to translate evidence-based interventions such as behavioural and exposure therapies to computer game formats (38). If game environments can be shown to facilitate changes in behaviours it

follows that changes in behaviour can be captured and potentially used to identify and monitor patterns of behaviour related to disease onset and progression. Thus, a gaming platform where participants have high engagement and compliance may be a useful adjunct in early intervention programs.

Our simulations of changes in model properties and model parameters are designed to provide a proof-of-principle that active inference can be used for hypothesis testing in clinical populations. To illustrate these sorts of model-based predictions we appeal to general ideas in the literature regarding synaptic loss in ageing and ideas related to reward sensitivity in anhedonia. We do not aim to provide validation of these effects – but rather illustrate how data from a clinical cohort could be explored in the context of a principled theory of brain function. The rationale behind the comparison of 6 and 10 state agents to recapitulate game play from older and younger adults respectively is based on our earlier work on the free energy principle and ageing (39). Theory and data support the idea that synaptic loss over the lifespan may offer adaptive pruning, where simpler models instantiated in older brains are driven by top-down prior beliefs (30, 39). This in turn will make older brains more resilient to short term changes in environmental input and provide a more general purpose brain where ‘the gist’ of an environmental challenge is readily identified. This is in contradistinction to younger brains which may over learn unimportant details of the environment’s structure.

In simulating features of depression, we choose anhedonia where previous meta-analytic work had highlighted the importance of diminished reward sensitivity but intact learning from reward and punishment (12). It is important to note that we do not verify this feature of anhedonia but rather use it as a demonstration of how symptomatic labels in psychiatric illness may be mapped to model parameters. We find that we can simulate a small behavioural deficiency in playing DOOM, by flattening the prior belief structure. This is a close correlate of diminished reward sensitivity but casts the phenomenon in the future, not

the present. Of course, patients with depression do have a diminished optimism about the future (40) though here we aim to demonstrate that goal states represent not only desired outcomes but also believed outcomes. We therefore provide a testable hypothesis where diminished optimism observed in patients is related to diminished capacity to perceive alternative outcomes during decision making (41) – under our framework they are the same thing. In future work this could be tested in a population of people with anhedonia – for example by using estimated prior belief structures from our model to predict individual optimism scores.

Overall, we show that as a dual normative and process theory of the brain, active inference under the free energy principle may be used to reveal structure in behaviour and imaging markers in novel experimental settings, allowing clinicians and patients to gain a more comprehensive description at the algorithmic and mechanistic level, of mental illness. We also suggest that active inference may be a more sensitive model of mind as compared to more traditional reinforcement learning models in the literature.

Acknowledgments

MC is funded by the EPRSC DTP. KJF is funded by the Wellcome trust.

Disclosures

The authors report no biomedical financial interests or potential conflicts of interest.

References

1. Montague, P. R., R. J. Dolan, K. J. Friston and P. Dayan (2012). "Computational psychiatry." Trends in cognitive sciences **16**(1): 72-80.
2. Friston, K. J., A. D. Redish and J. A. Gordon (2017). "Computational nosology and precision psychiatry." Computational Psychiatry **1**: 2-23.
3. Fletcher, P. C. and C. D. Frith (2009). "Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia." Nature Reviews Neuroscience **10**(1): 48-58.
4. Paliwal, S., F. H. Petzschner, A. K. Schmitz, M. Tittgemeyer and K. E. Stephan (2014). "A model-based analysis of impulsivity using a slot-machine gambling paradigm." Frontiers in human neuroscience **8**: 428.
5. Pedersen, M. L., M. J. Frank and G. Biele (2017). "The drift diffusion model as the choice rule in reinforcement learning." Psychonomic bulletin & review **24**(4): 1234-1251.
6. Redish, A. D. (2004). "Addiction as a computational process gone awry." Science **306**(5703): 1944-1947.
7. Hyett, M. P., M. J. Breakspear, K. J. Friston, C. C. Guo and G. B. Parker (2015). "Disrupted effective connectivity of cortical systems supporting attention and interoception in melancholia." JAMA psychiatry **72**(4): 350-358.
8. Tognin, S., W. Pettersson-Yeo, I. Valli, C. Hutton, J. Woolley, P. Allen, P. McGuire and A. Mechelli (2014). "Using structural neuroimaging to make quantitative predictions of symptom progression in individuals at ultra-high risk for psychosis." Frontiers in psychiatry **4**: 187.
9. O'Doherty, J. P., A. Hampton and H. Kim (2007). "Model□based fMRI and its application to reward learning and decision making." Annals of the New York Academy of sciences **1104**(1): 35-53.
10. Daw, N. D., S. Kakade and P. Dayan (2002). "Opponent interactions between serotonin and dopamine." Neural Networks **15**(4): 603-616.
11. Guitart-Masip, M., Q. J. Huys, L. Fuentemilla, P. Dayan, E. Duzel and R. J. Dolan (2012). "Go and no-go learning in reward and punishment: interactions between affect and effect." Neuroimage **62**(1): 154-166.
12. Huys, Q. J., D. A. Pizzagalli, R. Bogdan and P. Dayan (2013). "Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis." Biology of mood & anxiety disorders **3**(1): 12.
13. Sonuga-Barke, E. J. (2003). "The dual pathway model of AD/HD: an elaboration of neuro-developmental characteristics." Neuroscience & Biobehavioral Reviews **27**(7): 593-604.
14. Iglesias, S., C. Mathys, K. H. Brodersen, L. Kasper, M. Piccirelli, H. E. den Ouden and K. E. Stephan (2013). "Hierarchical prediction errors in midbrain and basal forebrain during sensory learning." Neuron **80**(2): 519-530.
15. Ahn, W.-Y., A. Krawitz, W. Kim, J. R. Busemeyer and J. W. Brown (2011). "A model-based fMRI analysis with hierarchical Bayesian parameter estimation." Journal of neuroscience, psychology, and economics **4**(2): 95.
16. Friston, K., T. FitzGerald, F. Rigoli, P. Schwartenbeck and G. Pezzulo (2017). "Active Inference: A Process Theory." Neural Comput **29**(1): 1-49.
17. Friston, K. and S. Kiebel (2009). "Predictive coding under the free-energy principle." Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1521): 1211-1221.
18. Bastos, A. M., W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries and K. J. Friston (2012). "Canonical microcircuits for predictive coding." Neuron **76**(4): 695-711.

19. Moran, Rosalyn J., et al. "Free energy, precision and learning: the role of cholinergic neuromodulation." Journal of Neuroscience 33.19 (2013): 8227-8236.
20. Flagel, S. B., et al. "A novel framework for improving psychiatric diagnostic nosology." (2016): 169-199.
21. Schwartenbeck, P., T. H. FitzGerald, C. Mathys, R. Dolan and K. Friston (2014). "The dopaminergic midbrain encodes the expected certainty about desired outcomes." Cerebral Cortex **25**(10): 3434-3445.
22. Schwartenbeck, P. and K. Friston (2016). "Computational phenotyping in psychiatry: a worked example." eneuro **3**(4): ENEURO. 0049-0016.2016.
23. Brockman, G., V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang and W. Zaremba (2016). "Openai gym." arXiv preprint arXiv:1606.01540.
24. Kempka, M., M. Wydmuch, G. Runc, J. Toczek and W. Jaśkowski (2016). Vizdoom: A doom-based ai research platform for visual reinforcement learning. Computational Intelligence and Games (CIG), 2016 IEEE Conference on, IEEE.
25. Mahendran, A., H. Bilen, J. F. Henriques and A. Vedaldi (2016). "ResearchDOOM and CocoDOOM: learning computer vision with games." arXiv preprint arXiv:1610.02431.
26. Gard, D. E., A. M. Kring, M. G. Gard, W. P. Horan and M. F. Green (2007). "Anhedonia in schizophrenia: distinctions between anticipatory and consummatory pleasure." Schizophrenia research **93**(1): 253-260.
27. Treadway, M. T. and D. H. Zald (2011). "Reconsidering anhedonia in depression: lessons from translational neuroscience." Neuroscience & Biobehavioral Reviews **35**(3): 537-555.
28. DeLisi, L. E. (1997). "Is schizophrenia a lifetime disorder of brain plasticity, growth and aging?" Schizophrenia research **23**(2): 119-129.
29. Jeste, D., E. Twamley, L. Eyler Zorrilla, S. Golshan, T. Patterson and B. Palmer (2003). "Aging and outcome in schizophrenia." Acta Psychiatrica Scandinavica **107**(5): 336-343.
30. Moran, R. J., M. Symmonds, R. J. Dolan and K. J. Friston (2014). "The brain ages optimally to model its environment: evidence from sensory learning over the adult lifespan." PLoS computational biology **10**(1): e1003422.
31. Friston, K., J. Mattout and J. Kilner (2011). "Action understanding and active inference." Biological cybernetics **104**(1-2): 137-160.
32. Friston, K., T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O'Doherty and G. Pezzulo (2016). "Active inference and learning." Neurosci Biobehav Rev **68**: 862-879.
33. Williams, J. and P. Dayan (2005). "Dopamine, Learning, and Impulsivity: ABiological Account of Attention-Deficit/Hyperactivity Disorder." Journal of Child & Adolescent Psychopharmacology **15**(2): 160-179.
34. Friston, K. J., T. Parr and B. de Vries (2017). "The graphical brain: Belief propagation and active inference." Network Neuroscience **0**(0): 1-34.
35. Cevasco, A. M., R. Kennedy and N. R. Generally (2005). "Comparison of movement-to-music, rhythm activities, and competitive games on depression, stress, anxiety, and anger of females in substance abuse rehabilitation." Journal of music therapy **42**(1): 64-80.
36. Burke, C. J., P. N. Tobler, M. Baddeley and W. Schultz (2010). "Neural mechanisms of observational learning." Proceedings of the National Academy of Sciences **107**(32): 14431-14436.
37. Saposnik, G., R. Teasell, M. Mamdani, J. Hall, W. McIlroy, D. Cheung, K. E. Thorpe, L. G. Cohen and M. Bayley (2010). "Effectiveness of virtual reality using Wii gaming technology in stroke rehabilitation: a pilot randomized clinical trial and proof of principle." Stroke **41**(7): 1477-1484.
38. Hudlicka, E. (2016). Virtual affective agents and therapeutic games. Artificial Intelligence in Behavioral and Mental Health Care, Elsevier**:** 81-115.

39. Gilbert, J. R. and R. J. Moran (2016). "Inputs to prefrontal cortex support visual recognition in the aging brain." Scientific reports **6**: 31943.
40. Sharot, T. (2011). "The optimism bias." Current Biology **21**(23): R941-R945.
41. Ambady, N. and H. M. Gray (2002). "On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments." Journal of personality and social psychology **83**(4): 947.
42. Harris, C. and M. Stephens (1988). A combined corner and edge detector. Alvey vision conference, Citeseer.

Figure Legends

Figure 1. Feature Extraction

Observation from the gym DOOM environment in 480x640 pixel space corresponding to state 1 (top). This observation is cropped to 100x640 pixels, removing image features such as the ceiling and game information to allow more efficient processing of the pixel data (top-middle). The location of the target in this space is determined by the Harris Corner detection operator (42), under the assumption that the target is present within the frame that exhibits the largest variation in (global) pixel intensity. Output from Harris Corner detection algorithm with local maxima of the corner response function highlighted in yellow (bottom-middle). We defined discrete 'states' of the environment by the location of the target (monster) relative to the player and whether the agent is currently shooting.

Figure 2. Adaptive Behaviours and Learned Contingencies

a) An overlay of the positional states for both 6 and 10 state environments, note that the size of the centre state is constrained to the size of the target such that the fire action remains effective. b) Simulated agents underwent a single trial of learning. The *B* matrices shown here correspond to the 'Fire', 'Move Right', 'Move Left' actions at t=4, t=16, t=64 and t=128 under the free energy minimization (upper) and reward-maximizing (lower) paradigms. Each matrix represents the agent's belief about how the environment will change after making the respective action. For example, at trial 4 the Reinforcement Learning agent strongly believes that a 'fire' action will bring it from state 3 to state 4. The uniformity of the 'move right' matrix at the same time step implies that the agent has no knowledge of the consequence of a 'Move Right' action. After 128 epochs of learning, the B matrices of the free energy agent have converged to those of the optimized agent presented in Figure 3A. The B matrix of the

reward-based agent is much sparser by comparison, reflecting a lack of knowledge about the environmental contingencies.

c) Reward metrics collected from the free energy (upper) and reward maximizing (lower) agents in B.

Figure 3. Game Structure and Model Comparisons

A graphical representation of the possible state transitions within the 6 (left) and 10 (right) state environments. Green lines denote optimal transitions from each state while red arrows denote possible but sub optimal transitions. Any connection not shown is not possible within the DOOM environment, for example, it is not possible to move from a 'right and firing' (**RF**) state to a 'middle and firing' (**MF**) state without transitioning through the 'middle not firing' (**MNF**) state.

Figure 4. Comparison of Free Energy and Reward Maximizing Agents

Reward (left) and survival (right) metrics collected from 50 free energy (blue) and reward-maximizing (black) agents. The mean total reward achieved by the free energy agents was significantly greater than that of the reward maximizing agent; $p = 0.05$. Plot shows mean +/- s.e.m.

Figure 5. Comparison of Free Energy Agents and Human Players

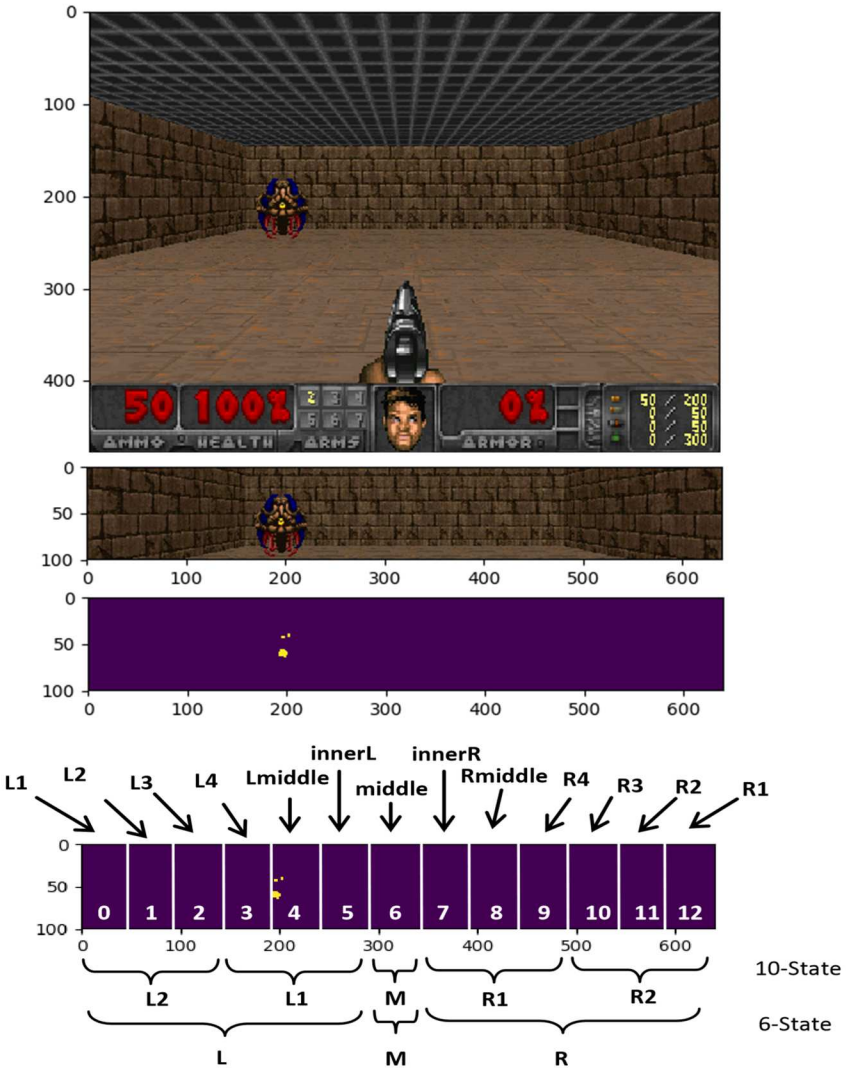
Human participants played 64 episodes of the DOOM game. The upper panel shows the average survival scores (lower is better) +/- s.e.m. These were compared to the free energy agents from figure 2 – where alternating trials from the 128 episodes were compared to the human's 64 episodes. The lower panel shows a 'Manhattan plot', of statistical difference (-log(p-value)) for each episode. After 6 epochs (12 decisions) the free energy agents attain human-like performance.

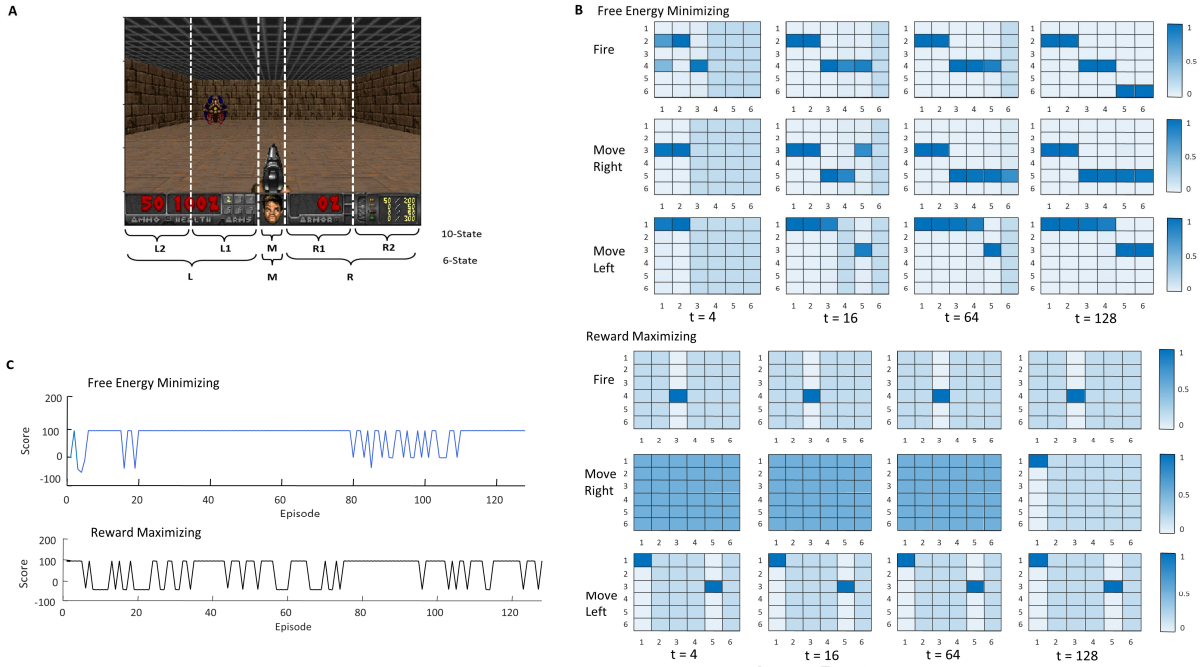
Figure 6. Simulations of Ageing

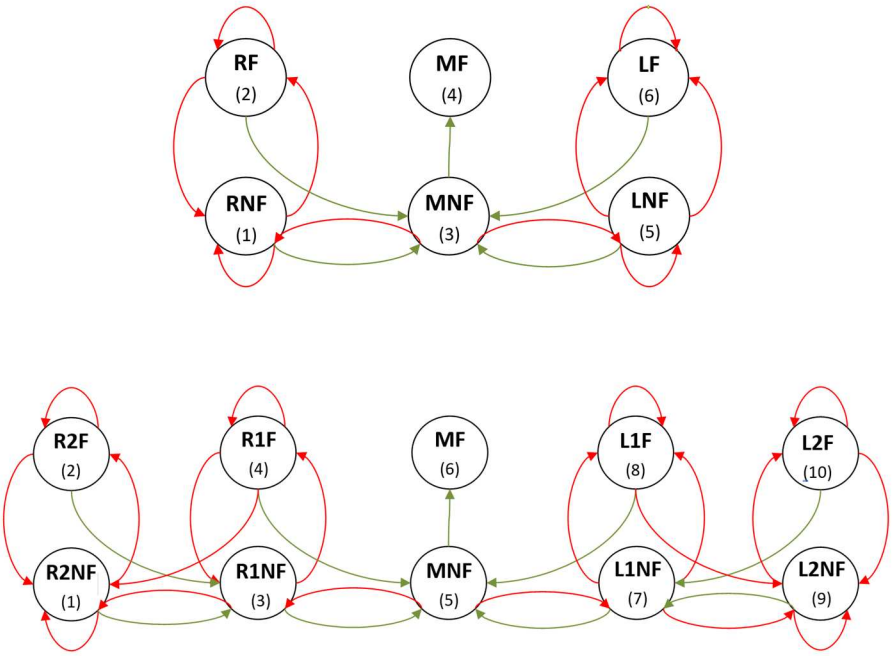
a) Comparison of 6-state (black) and 10-state (red) free energy agents shown again for every other episode over 128 episodes. Plots denote mean +/- s.e.m. A quadratic polynomial fit to each average curve (6 state and 10 state) is superimposed to illustrate qualitative similarities with the human age effects. b) Comparison with ageing effects in human play. Similar to the simulated agents, older participants show shallower progression in the game over many episodes (Plots denote mean +/- s.e.m. with a superimposed second order polynomial fit to the mean). Linear trends in the data were more strongly negative for the younger cohort and the more complex state space model.

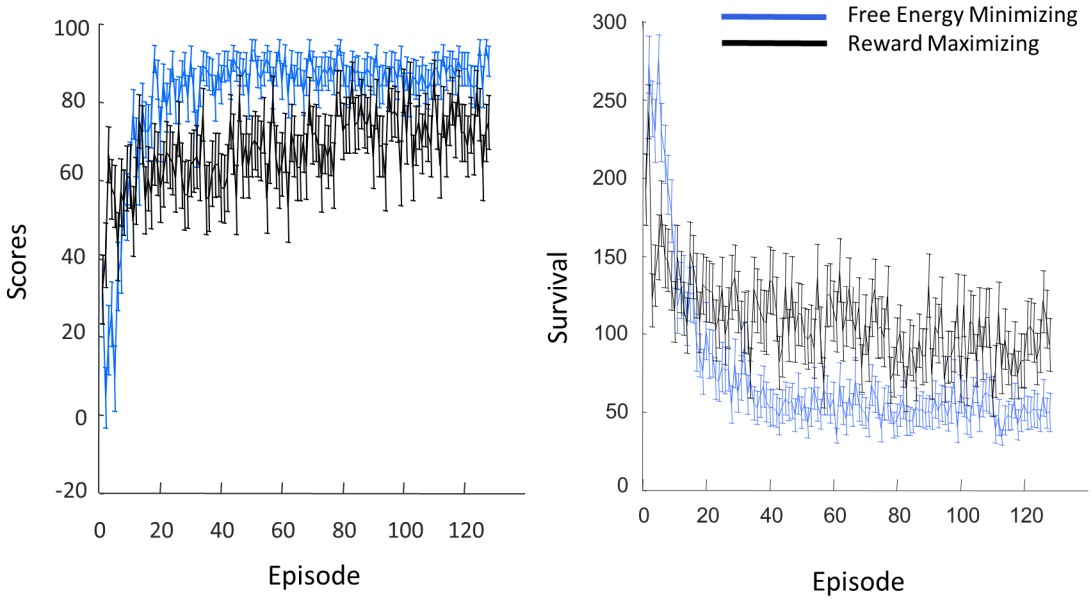
Figure 7. Simulations of Anhedonia

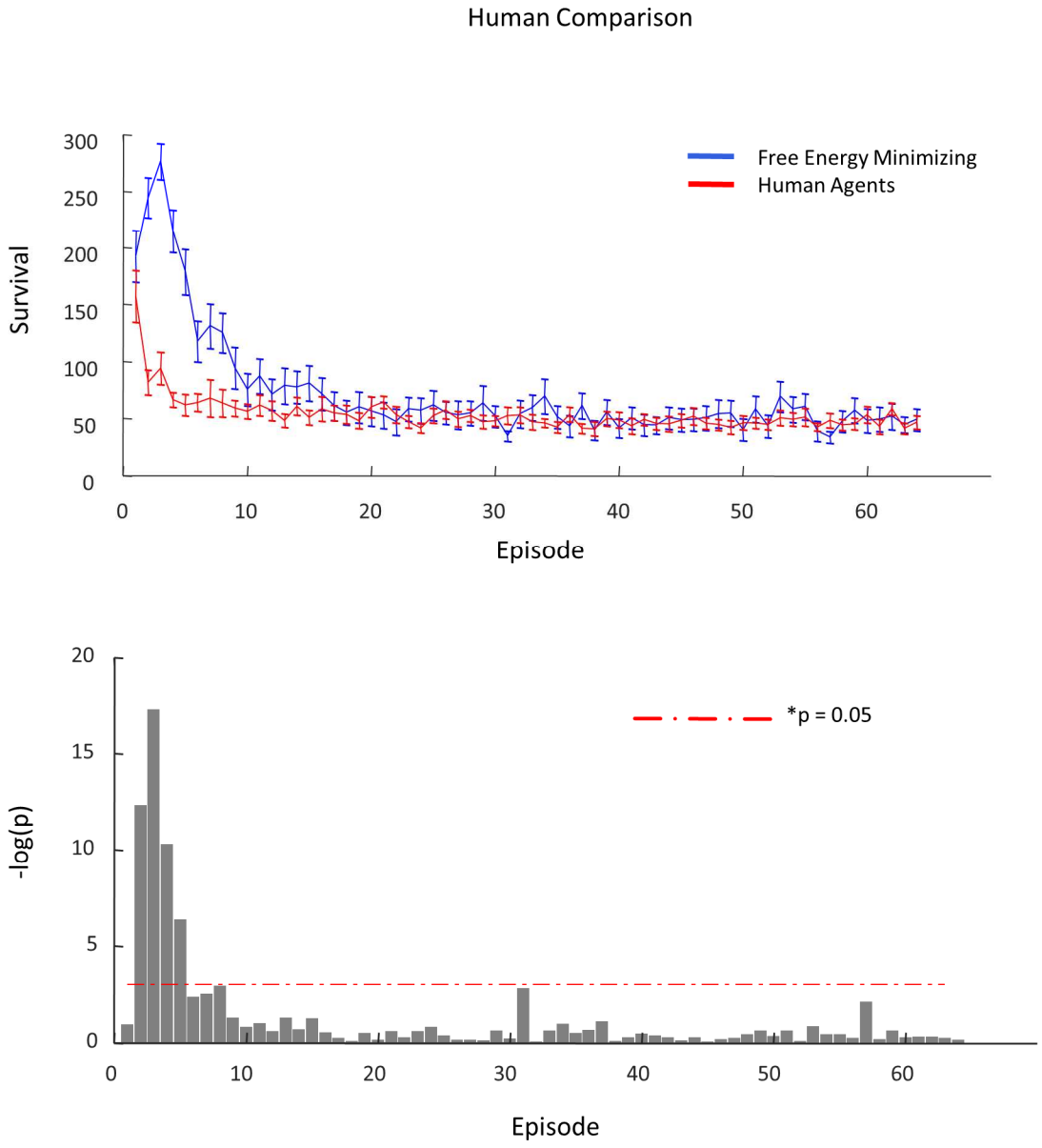
a) Prior belief structure for the 'motivated': green compared to the 'anhedonic': blue agents who carry a 10 state model. The anhedonic agent displays a flattened prior belief in the final state of the game, believing with less magnitude that it will kill the monster compared to the motivated agent. Behavioural performance was significantly worse for the anhedonic agents ($p < 0.05$); however, performance matched the motivated agents later across the 64 trials. Example of imposed trial timings for simulated agents. B) Local field potentials derived from state updates that evaluate previous current and future states under all allowable policies. Plotted LFPs are proposed to thus represent the prefrontal cortex. When comparing a single motivated to anhedonic agent the LFPs exhibit large differences around trial 20 and persist over ~5 trials. This finding was replicated in 3 other agent comparisons. c) Based on these LFPs we simulated the associated BOLD response from the PFC and show a second increase in the HRF around 20 seconds for the motivated compared to the anhedonic agents. Anhedonic agents exhibit a more protracted HRF.

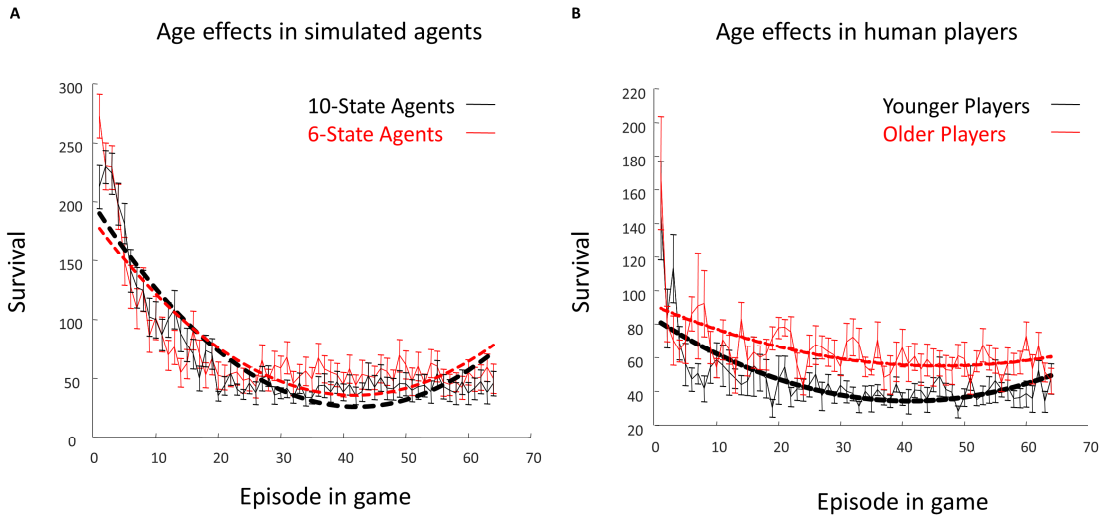


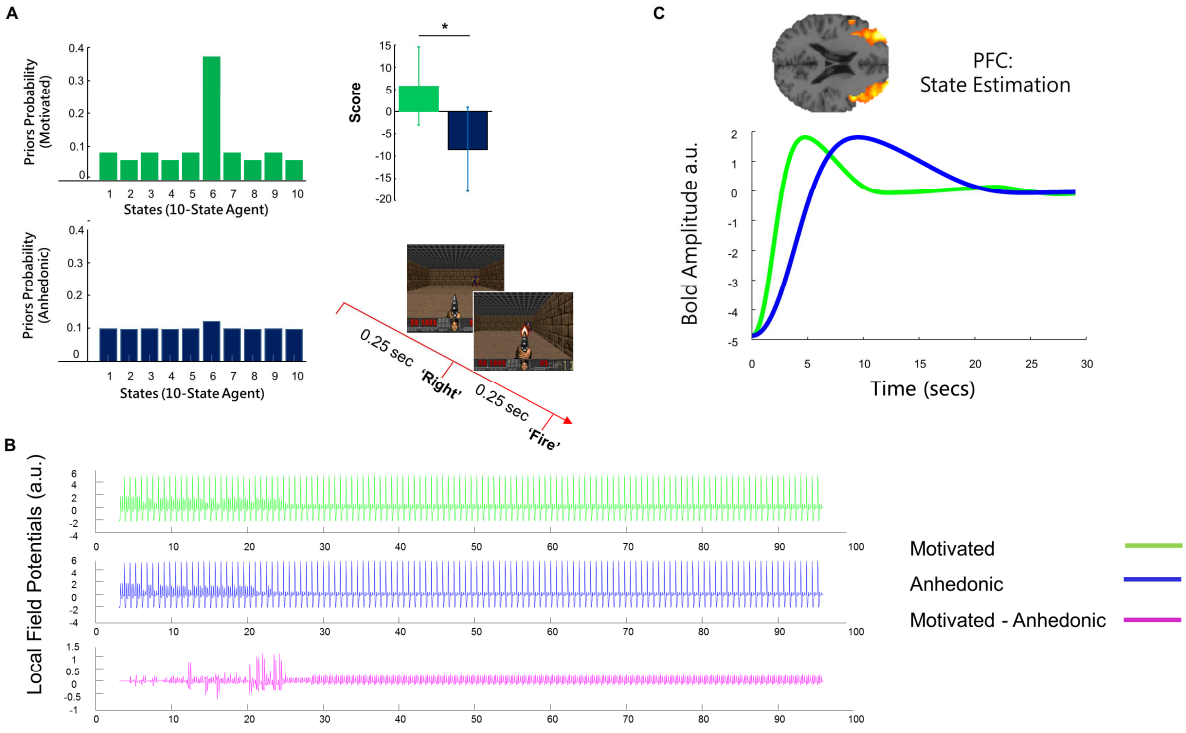












Active Inference on OpenAI Gym: A Paradigm for Computational Investigations into Psychiatric Illness

Supplemental Information

Active Inference and Reinforcement Learning

Active inference rests on a generative model of observed outcomes that can be optimized with respect to their variational free energy, an information theoretic construct representing the relative entropy between the true states of the world and models (of the world) in the brain. Recent formulations of decision making assume a partially observable Markov Decision Process (POMDP) form of the generative model (1, 2, 3, 4). This model defines the joint probability distribution over the observations, hidden states, policies (a sequence of available actions) and precision (or degree of belief in controllability of the environment). A graphical representation of the generative model is illustrated in Supplementary Figure S1. Here, a likelihood term establishes a mapping between hidden states and observations, defining the probability of being in a state after making an observation. This is referred to throughout this paper as the *A* or observation matrix. For our game play we assumed an identity matrix for *A* and leave uncertainty about hidden states to emerge in the transition probabilities. A hidden state may be thought of as the state of an entity or an environmental property that may not be directly observed by the agent; within the DOOM environment this corresponds to the relative position of the agent to the monster. This is because our feature extraction is imperfect, hence the agent cannot directly observe its position relative to the monster and the states are uncertain or 'hidden'. The probability of transitioning from one hidden state to another is encoded within the *B,* or transition matrix. For example, if action *k* submitted from positional state *i* will move the agent to state *j* with absolute certainty, this will be represented in matrix form as B*{k}*(*i*,j) = 1, where all other elements in the row vector *i* will be equal to 0. Supplemental Figure S2

illustrates potential trajectories within the DOOM environment in matrix form. For our simulations below, we set each entry in each column of *B* to have equal values, i.e. the agents did not know the true environmental contingencies and had to learn them.

Finally, the mapping between policies and hidden states are also influenced by the agent's prior preferences, which determine how likely or rewarding a given outcome is. This critical quantity is denoted as the *C* vector and profiles the prior preferences or 'end goal' of the game. In the DOOM environment it should be maximized at the position in front of the monster firing. Utility, or the agent's sense of reward is quite literally the absence of prediction error; 'The agent will find this outcome rewarding' and 'the agent believes this outcome is likely' are equivalent statements. The probability of a policy also depends on precision γ and its hyperparameters α and β. Precision is related to the inverse temperature parameter from statistical physics and softmax response functions economics (but is optimized with respect to free energy over play) and determines an agent's confidence in its decisions.

Based upon the current form of the generative model, an action is chosen from a particular policy (π) where that policy minimizes the expected free energy of the agent (Eqn. 1). To evaluate the expected free energy of a policy and select an action; i.e., whether to move left, move right or fire, the agent first must estimate its current, past and future hidden states under each available policy. In our simulations, we allow the agent to entertain short horizon policies (3-action sequences) and allow all combinations of such 3-action policies to give a total number of policies of 9 e.g. one policy may be {'left', 'left', 'left'}, while another {'right', 'right', 'fire'}. State estimation also minimizes free energy according to Supplementary Figure S1. After a set of 16 iterative updates to estimate the states under each policy, a Bayesian Model Averaging procedure (5) is used to construct the final expected states in the past, currently and in the future $q(s)$. Using this estimate, the (negative) expected Free Energy of each policy is

calculated and passed through a softmax operator to select the current best policy and hence the current optimal action: 'left', 'right' or 'fire'. Within the softmax operation the precision parameter determines the agent's confidence in the decision, which is itself updated at the end of each policy evaluation cycle (6).

Under active inference a policy, π at time t, is valuable (has a high negative expected free energy, $Q(\pi, t)$) when it maximises the expected information (7) about the true state of the environment (i.e. maximizing epistemic value – first term expected under current state estimate) while maximizing extrinsic value (reward of getting to the preferred or believed final state – second term expected under current state estimate).

$$Q(\pi|t) = \langle lnP(o|s)| - lnP(o|\pi)\rangle + \langle lnP(o)\rangle \qquad \text{Eqn. 1}$$

What emerges is an adaptive agent that moves—purposefully—through an environment to solicit outcomes that it believes are most likely (i.e., the least surprising). For example, an agent might believe that rewards are likely outcomes and therefore act to minimise surprise by maximizing reward. Exploratory behaviours are encouraged in a principled trade-off between epistemic information and explicit value or reward (Eqn 1). This is mathematically what the Free Energy cost function does - without needing to recourse to ad-hoc exploratory features. If there is uncertainty in the environment structure, the agent will, by definition, explore to some extent.

In the literature, the application of reward-based models to neural and behavioural data has previously evoked the difference between 'model-based' and 'model-free' reinforcement learning (8). Here the distinction is whether the transition matrices are explicitly learned through state prediction errors (model-based) or not (model-free). Thus, our reward-based and

active inference-based systems are both of the 'model-based kind, whereby both are allowed to learn state transitions.

Given this form it is easy to contrast active inference with reinforcement learning or simply 'reward maximising' agents. For this comparison (Free Energy Minimizing vs. Reward Maximising), we simply remove the epistemic value term from the evaluation of the policy.

To simulate DOOM play under active inference, we allow the agent to learn the optimal state transitions. Our learning scheme treats the model's transition probabilities (B) as unknown and establishes beliefs over these unknowns in the form of Dirichlet distributions. The B matrices are updated after an action-observation cycle by incrementing at the observed state-outcome index. In this way, an agent will believe a state-outcome combination to be likely if it observed frequently. Using a set of 'flat' priors (i.e. each element of the matrix is set to 0.167 for a six-state model and 0.1 for the ten-state model) enables us to establish how a simulated agent plays the game with no de-novo knowledge about the environment. As noted above the A matrix is set to the identity, while for the prior belief in final states, we set to C = [0.1, 0, 0.2, 0.6, 0.1, 0] for the six-state model and C = [0.05, 0, 0.05, 0, 0.2, 0.6, 0.05, 0, 0.05, 0]. These prior beliefs about preferred states might correspond to how the game would be described by a human player; i.e., seek to be in-front of the target (states 3 and 5 for the 6-state and 10-state agents) then fire to win, (states 4 and 6 for the 6-state and 10-state model), while non-firing states are preferable to wasting ammunition (e.g. the zero belief in firing when the monster is on the left or right; states 2 and 6 in the six-state model, Figure 1b). For each artificial agent we simulated learning over 128 episodes and performed 50 runs to evaluate the average behaviour of the agent.

The DOOM Game Environment

At the beginning of each game the agent starts in centre of the screen while the target can be positioned anywhere in front of the agent but constrained to the back wall. Each 'episode' is defined as the period between initialization of a game and the end of the game, triggered by the agent shooting the target or 10 seconds (350 frames) of game time elapsing. The agent can then move through the environment with left and right movements or can emit a shot, what we denote as a ‘fire’ action, i.e. there are three possible actions.

The game metrics used to evaluate the performance of an agent are based on the total amount of reward obtained during an episode and the amount of time it takes to solve the episode. When a ‘fire’ action is submitted from the middle position, killing the monster, a reward value of 100 is returned. A negative reward penalty (-1) is imposed for every time step the agent moves or submits a 'fire' action from one of the left or right positions, meaning the agent loses points for staying alive while not solving the environment. The optimal solution to the task is thus to move in front of the target and then fire, resulting in a low number of steps and high reward score. In addition, the game returns a ‘survival’ score which serves to index the time taken for the monster to be killed.

To ensure that no agent performs better due to chance (starting closer to the target state), the environment creation processes has been seeded such that each agent is presented with the same series of environments.

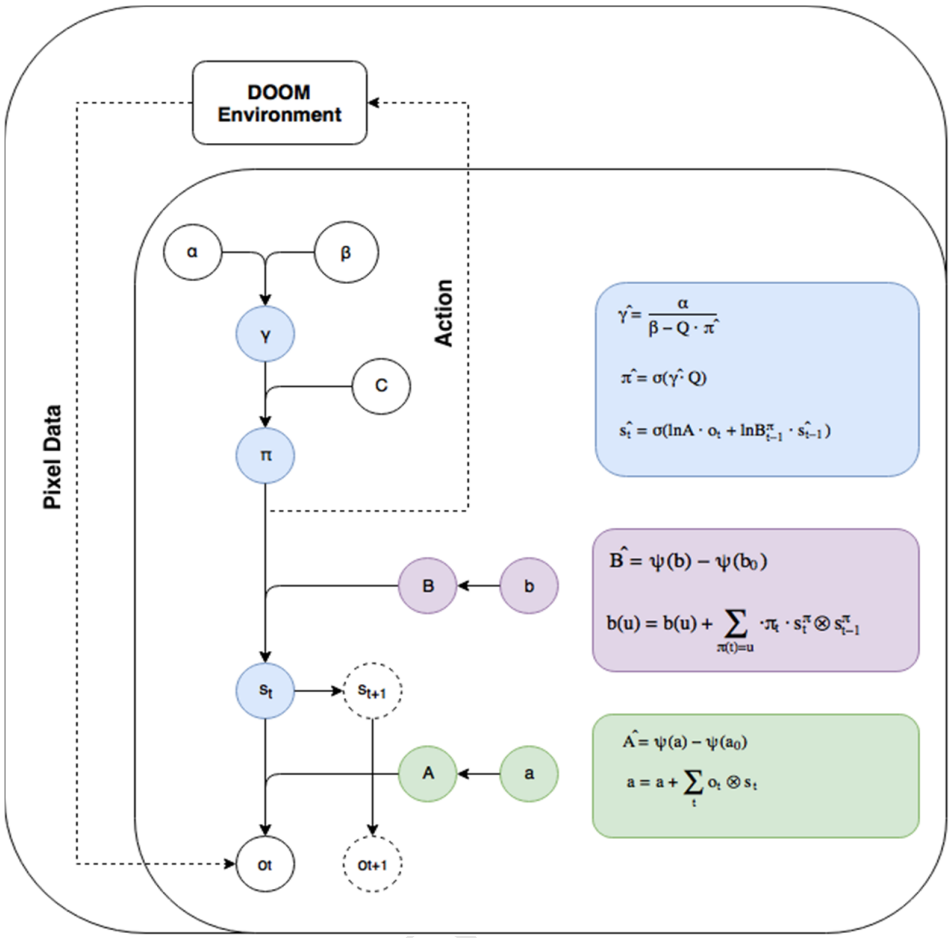
Human Players playing DOOM

We recruited two groups of players comprising younger (n=9, 22 ± 1 years) and older (n=7, 56 ± 5 years) adults. Subjects were recruited from the campus of the University of Bristol – from students and staff. Screening of subjects relied on self-report of no psychiatric or neurological history via a participant information sheet. Each participant played 64 consecutive episodes where each episode consisted of self-timed button-press responses and ended either when the monster was killed or when the game timed out (though in practice no human player was timed out). The players were told that the goal was to shoot the monster whilst conserving ammunition and each episode resulted in a survival score commensurate with the simulated agent's play. The buttons were 'left' 'right' and 'space' for all participants however the mapping in terms of move left, right or shoot was scrambled for each participant so they had to learn the correct button associated with each of the three possible game actions. While humans will obviously understand the general structure of the task (i.e. have ready access to transition matrices once a few buttons are depressed), we added this component to introduce a minor and rapid learning phase (Figure 4). This was done to mimic some form of the unknown action-related state transitions within the learning simulations of the *in silico* agents.

Ethics was approved by the Faculty of Biomedical Sciences Research Ethics Committee (FREC) at the University of Bristol.

Simulating Neuronal Responses

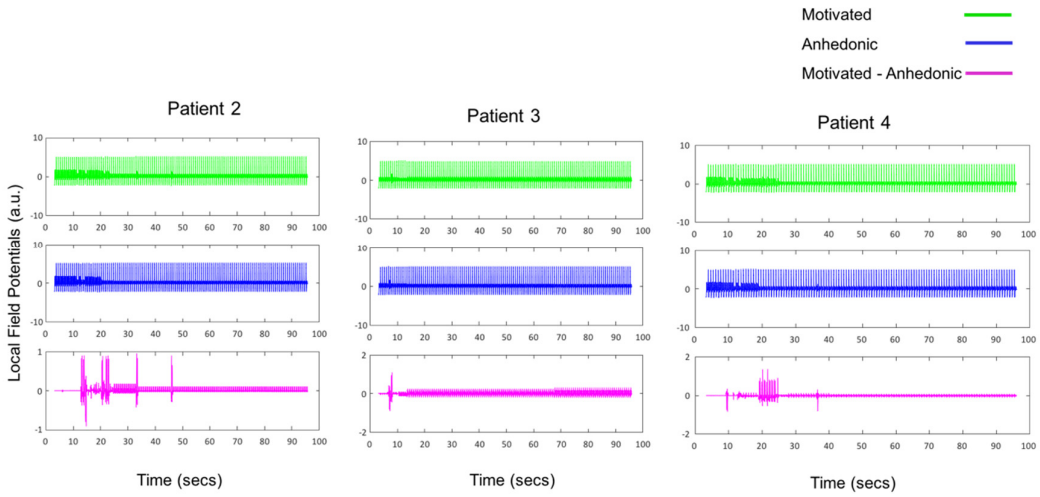
In order to examine the putative neurobiological correlates of free energy minimization over game play we performed the following steps: 1. Before an action is selected, the agent will estimate its current state and future states based upon previous observations and the expected action-dependent state transitions. 2. This procedure comprises a state estimation scheme that uses iterative gradients to get a 'best' estimate of the current and future states. An implementation of this scheme ('spm_MDP_VB.m') may be found in the DEM toolbox of the academic software SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm). 3. We assume the brain is doing this gradient descent. 4. Thus we take each gradient based estimate and assume that these fluctuations in estimates are directly mappable to firing in the prefrontal cortex 4. In order to simulate an LFP that may emerge from this computation we simple plot these estimates after a bandpass filtering procedure. 5. In order to do this we must assume a time associated with each computation – we choose a working time update of 16 msec. 6. Finally, to generate an associated BOLD response we pass the simulated LFP through a second function ('spm_hfx_hdm.m' also in the spm software package) based on a hemodynamic response function. This is a model of neurovascular coupling that maps local electrical brain activity to changes observed using fMRI (9).



Supplementary Figure S1. Separation between the DOOM environment (outer) and the agent's generative model (inner), formulated as an MDP. The agent is limited to one of three actions at a given time step (the action space is simply 'move right', 'move left', or 'fire') and. The 'state space' describes all possible states of the agent in the environment following the decision to emit an action. The state space in our simple 6-state DOOM models includes the states 'left-of-monster, not firing', 'left-of-monster, firing', 'right-of-monster, not firing', 'right-of-monster, firing', 'centered-on-monster, not firing' and 'centered-on-monster, firing'. The number of states the agent can occupy is thus relatively small, with a single stationary target state ('centered-on-monster, firing'). The inner figure demonstrates how the state transition matrix *B*, observation matrix *A* and prior expectations *C* influence action selection, belief updating and learning. Policy selection depends on the agent's prior preferences *C* and the prescribed precision **γ.** The *B* matrix provides a mapping between hidden states under the given action. The *A* matrix maps hidden states to observations and defines the probability of being in a given state after receiving an observation. Learning equates to updating *B* by accumulating evidence for real state transitions in consequence of action.

$$B_{fire} = \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&1&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1 \end{bmatrix} \quad B_{left} = \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix} \quad B_{right} = \begin{bmatrix} 1&1&1&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&1&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

Supplementary Figure S2. Composition of 10-state transition matrices that reflecting accurate beliefs about the contingencies of the DOOM environment.



Supplementary Figure S3. Reproductions of 'LFP responses' generated by our model via state estimate updates during game play from an 'anhedonic' and a 'motivated' agent. We show that the motivated agent displays larger LFP responses centered around trial 20 which could be observed as BOLD differences using fMRI (see main text Figure 5 for another exemplar trace).

Supplemental References

1. Friston, K., T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O'Doherty and G. Pezzulo (2016). "Active inference and learning." Neurosci Biobehav Rev 68: 862-879.

2. Pezzulo, G., E. Cartoni, F. Rigoli, L. Pio-Lopez and K. Friston (2016). "Active Inference, epistemic value, and vicarious trial and error." Learning & Memory 23(7): 322-338.

3. Friston, K., T. FitzGerald, F. Rigoli, P. Schwartenbeck and G. Pezzulo (2017). "Active Inference: A Process Theory." Neural Comput 29(1): 1-49.

4. Friston, K. J., M. Lin, C. D. Frith, G. Pezzulo, J. A. Hobson and S. Ondobaka (2017). "Active Inference, Curiosity and Insight." Neural Comput: 1-51.

5. Stephan, Klaas Enno, et al. "Bayesian model selection for group studies." *Neuroimage* 46.4 (2009): 1004-1017.

6. Friston, K., P. Schwartenbeck, T. FitzGerald, M. Moutoussis, T. Behrens and R. J. Dolan (2014). "The anatomy of choice: dopamine and decision-making." Phil. Trans. R. Soc. B 369(1655): 20130481.

7. Shewry, M. C. and H. P. Wynn (1987). "Maximum entropy sampling." Journal of Applied Statistics 14(2): 165-170.

8. Gläscher, Jan, et al. "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning." *Neuron* 66.4 (2010): 585-595.

9. Buxton, Richard B., Eric C. Wong, and Lawrence R. Frank. "Dynamics of blood flow and oxygenation changes during brain activation: the balloon model." *Magnetic resonance in medicine* 39.6 (1998): 855-864.